<table>
<thead>
<tr>
<th>Title</th>
<th>Variational Bayesian Sparse Signal Recovery With LSM Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Zhang, Shuanghui; Liu, Yongxiang; Li, Xiang; Bi, Guoan</td>
</tr>
<tr>
<td>Date</td>
<td>2017</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/44225">http://hdl.handle.net/10220/44225</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2017 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See <a href="http://www.ieee.org/publications_standards/publications/rights/index.html">http://www.ieee.org/publications_standards/publications/rights/index.html</a> for more information.</td>
</tr>
</tbody>
</table>
Variational Bayesian Sparse Signal Recovery With LSM Prior

SHUANGHUI ZHANG, YONGXIANG LIU, XIANG LI, AND GUOAN BI

1School of Electronic Science, National University of Defense Technology, Changsha 410073, China
2School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Corresponding author: Yongxiang Liu (lyx_bible@sina.com)

This work was supported in part by the National Science Foundation of China under Grant 61422114 and in part by the Natural Science Fund for Distinguished Young Scholars of Hunan Province under Grant 2015JJ1003.

ABSTRACT This paper presents a new sparse signal recovery algorithm using variational Bayesian inference based on the Laplace approximation. The sparse signal is modeled as the Laplacian scale mixture (LSM) prior. The Bayesian inference with the Laplacian models is a challenge because the Laplacian prior is not conjugate to the Gaussian likelihood. To solve this problem, we first introduce the inverse-gamma prior, which is conjugate to the Laplacian prior, to model the distinctive scaling parameters of the Laplacian priors. Then the posterior of the sparse signal, approximated by the Laplace approximation, is found to be Gaussian distributed with the expectation being the result of maximum a posterior (MAP) estimation. Finally the expectation-maximization (EM)-based variational Bayesian (VB) inference is utilized to accomplish the sparse signal recovery with the LSM prior. Since the proposed algorithm is a full Bayesian inference based on the MAP estimation, it achieves both the ability of avoiding structural error from the sparse Bayesian learning and the robustness to noise from the MAP estimation. Analysis on experimental results based on both simulated and measured data indicates that the proposed algorithm achieves the state-of-art performance in terms of sparse representation and de-noising.

INDEX TERMS Sparse signal recovery, Laplacian scale mixture (LSM), the Laplace approximation, expectation-maximization (EM), variational Bayesian inference.

I. INTRODUCTION

Sparse signal recovery or compressive sensing (CS) has attracted extensive research attention due to its desirable capacity of performance improvement for many practical applications. It is generally expressed as a linear regression problem in which the sparse signal is projected into a low dimensional space via

\[ y = \Phi w + n \]  

where \( y \in \mathbb{R}^N \), \( \Phi \in \mathbb{R}^{N \times M} \), \( w \in \mathbb{R}^M \) and \( n \in \mathbb{R}^N \) represent the observation, the measurement matrix, the sparse signal and the noise, respectively. Because recovering \( w \) from \( y \) is ill posed for \( N < M \), additional information is therefore needed to solve the linear regression problem in (1). For sparse signal recovery, sparse prior of the signal is generally utilized to solve this ill posed problem. In general, the \( l_0 \) norm of the sparse signal denotes the number of the non-zero elements and is used to represent the sparse degree. Therefore, (1) can be solved under the \( l_0 \) norm constraint as

\[ \hat{w} = \arg \min_w \left\{ \|y - \Phi w\|_2^2 + \tau \|w\|_0 \right\} \]  

where \( \| \cdot \|_k \) represents the k-norm operator and \( \tau \) is the sparse constraint coefficient. It has been reported that the solution to the above equation is NP-hard [1]. Therefore, other alternative cost function, e.g. \( l_1 \) and \( l_p \) norm, are introduced for better tractability.

Most sparse signal recovery algorithms are based on the \( l_1 \) or \( l_p \) norm, such as the Basis Pursuit [2], the iterative reweighted \( l_1 \) and \( l_2 \) methods [3], [4], the smoothed \( l_0 \) norm method [5], and so on. Greedy pursuit is another widely used strategy in sparse signal recovery, which generally acquires high computational efficiency with low estimating precision. Representative greedy method is the Orthogonal Matching Pursuit (OMP) [6], [7].

Sparse signal recovery problem can also be solved within the Bayesian framework. Compared with the norm regularized methods, the Bayesian sparse signal recovery
algorithms show better performances in many respects. The model parameters can be learned automatically within the Bayesian framework to avoid manually adjusting of the parameters. The Bayesian methods can find the global minima more easily than the sparse regularized methods. Finally additional statistical information of the sparse signal can be easily utilized with a suitable prior. The maximum a posterior (MAP) estimation and the hierarchical Bayesian estimation are two main strategies used in Bayesian sparse signal recovery. For MAP estimation, the sparse resolution is located at the peak of the posterior, and the general analytical form of the posterior can be avoided. It has been found that MAP estimation is equivalent to the norm regularized method by choosing suitable prior. For example, the Laplacian prior based MAP estimation induces the norm minimization problem such as LASSO [8]. The MAP estimation based Bayesian sparse signal recovery suffers from a large probability of local minima, since it is essentially a penalized regression and does not use any statistical information of the posterior [9]. In contrast, the hierarchical Bayesian framework is a full Bayesian inference, wherein the sparse signal is modeled by some hierarchical priors, and estimated by the expectation of the posterior. The most well-known hierarchical Bayesian method is the Sparse Bayesian Learning (SBL) [10]–[12]. It uses the hierarchical Gaussian prior with the inverse variance being gamma distributed to model the sparse signal. It has been proved that SBL obtains fewer local minima than the MAP estimation and avoids the structural error. In recent years, SBL has been modified to fit the block sparsity problems [13] and the multiple measurement vectors problems [14].

In general, the Laplacian prior is more suitable to be used to represent sparse signals than the Gaussian prior, since its probability density function (PDF) has narrower pulse width with higher tail values than the Gaussian prior [15]. However, the Laplacian distribution is not conjugate to the Gaussian likelihood so that full Bayesian inference cannot be accomplished when the Laplacian prior is introduced [17]. Some tricks should be used to make the sparse Bayesian inference with the Laplacian prior possible. In [15], the hierarchical Gaussian prior is introduced to represent the Laplacian prior for the Bayesian compressive sensing based on Laplacian prior (BCSL). Essentially speaking, however, it is still in a gaussian scale mixture (GSM) form. Laplacian scale mixture (LSM) prior is used to model the sparse signal in [16] by only using MAP estimation to find the solution, which suffers from more local minima and structural error.

In this paper, we propose a sparse signal recovery algorithm using the variational Bayesian inference based on the Laplace estimation with LSM prior. Instead of the MAP estimation used in [16], full Bayesian reference with LSM prior is designed by addressing the following issues. The first one is to propose the scaling parameter of Laplacian prior to be inverse-gamma distributed for conjugating to Laplacian distribution. The second one is introducing the Laplace approximation [21] to estimate the posterior of the sparse signal by deriving a second order Taylor expansion of the posterior around the result of MAP estimation, which allows the posterior to be approximately Gaussian distributed with the expectation being the result of MAP estimation. Once the posterior is derived, the variational Bayesian method based on the expectation-maximization (EM) is used to accomplish sparse signal recovery with the LSM prior. Experimental results show that the proposed algorithm achieves both the ability of avoiding structural error from SBL and the robustness to noise from the MAP estimation.

This paper is organized as follows. The LSM model for the sparse representation is presented in section II and the approximate variational Bayesian inference for LSM prior is derived in section III. The performances of the proposed method and SBL, BCSL, OMP are compared by using simulated and measured data in section IV. Finally, the paper is summarized in section V.

II. SPARSE SIGNAL MODEL WITH THE LSM PRIOR

The graphical model of LSM prior is shown as Fig. 1, where the squared nodes correspond to parameters of the model, the nodes with double circles represent observed random variables and the nodes with single circle represent latent random variables.

The observation model is shown as (1). The additive noise, \( n \), is commonly assumed to be the white Gaussian noise, which is independent, zero mean, and Gaussian distributed:

\[
p (n) = \mathcal{N}(n | 0, \beta^{-1} I)
\]

where \( \beta \) is the noise precision or reciprocal of the variance, and \( I \in \mathbb{R}^{N \times N} \) is an identity matrix. Since the Gamma distribution is conjugate to the Gaussian distribution, the inverse variance \( \beta \) is generally supposed to be Gamma distributed [17]:

\[
p (\beta; a, b) = \mathcal{G}(\beta; a, b) = \frac{b^a}{\Gamma(a)} \beta^{a-1} \exp(-b\beta)
\]

where \( \Gamma(a) \) is the Gamma function, and the scaling parameters \( a \) and \( b \) are generally fixed to small values to make the prior of \( \beta \) non-informative, e.g. \( a = b = 10^{-4} \) [10].

\[
\Gamma(a) = \int_0^{+\infty} x^{a-1} \exp(-x) dx
\]
The likelihood of the observation \( y \) can be derived from (1) as:

\[
p(y|w, \beta) = N\left(y|\Phi w, \beta^{-1}I\right)
\]

\[
= (2\pi)^{-\frac{N}{2}} \beta^{-\frac{N}{2}} \exp\left(-\frac{\beta}{2} \|y - \Phi w\|^2\right).
\]

In Bayesian framework, the sparse signal, \( w \), is usually assumed to be random variables and modeled with some sparse priors. The widely applied sparse prior is the GSM prior, which is a Gaussian prior with the inverse variance being Gamma distributed, and is convenient to conduct the posterior. In contrast, this paper introduces the LSM prior to be Gamma distributed, and is convenient to conduct the assumption to be random variables and modeled with some Laplacian distributions with distinctive scaling parameters, \( \lambda \).

\[
\lambda
\]

<table>
<thead>
<tr>
<th>( \lambda )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplacian distribution</td>
</tr>
<tr>
<td>student-t distribution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \lambda )</th>
</tr>
</thead>
<tbody>
<tr>
<td>distribution</td>
</tr>
<tr>
<td>GPD</td>
</tr>
</tbody>
</table>

\[
\text{Fig. 2, for (5) and (7), the marginal likelihood of } w \text{ achieved to be}
\]

\[
\lambda
\]

\[
\text{ posterior.}
\]

\[
\text{In order to flexibly model local characteristics of the sparse coefficients, a set of nonstationary Laplacian distributions with distinctive scaling parameters, } \lambda_m (m = 1, \cdots, M), \text{ are used to model them:}
\]

\[
p(w|\lambda) = \prod_{m=1}^{M} L(w_m|0, \lambda_m)
\]

\[
= \prod_{m=1}^{M} \frac{1}{2\lambda_m} \exp\left(-\frac{|w_m|}{\lambda_m}\right). \quad (6)
\]

Noting that inverse-Gamma distribution is conjugate to the Laplacian distribution, the scaling parameters, \( \lambda_m \), are supposed to be inverse-Gamma distributed:

\[
p(\lambda; c, d) = \prod_{m=1}^{M} IG(\lambda_m; c, d)
\]

\[
= \prod_{m=1}^{M} \frac{d^c}{\Gamma(c)} \lambda_m^{c-1} \exp\left(-\frac{d}{\lambda_m}\right). \quad (7)
\]

Also, to make the prior non-informative, the scaling parameters \( c \) and \( d \) are set to be small, e.g., \( c = d = 10^{-4} \).

Since the factorial model of the sparse signal \( w \) is given in (6) and (7), the marginal likelihood of \( w \) can be analytically achieved to be

\[
p(w; c, d) = \prod_{m=1}^{M} p(w_m|c, d)
\]

\[
= \prod_{m=1}^{M} \int_{0}^{+\infty} p(w_m|\lambda_m) p(\lambda_m; c, d) d\lambda_m
\]

\[
= \prod_{m=1}^{M} \frac{cd^c}{2(|w_m| + d)^{c+1}} \quad (8)
\]

by eliminating the scaling parameter \( \lambda_m \), as shown in Appendix A.

It is interesting to see that the expression in (8) is coincident with the generalized Pareto distribution (GPD) [18]. In order to observe the feasibility of LSM being used as a sparse prior, the PDFs of Laplacian distribution, student-t distribution (i.e., the marginal distribution of GSM) and GPD are compared in Fig. 2, for \( a = b = c = d = \lambda = 1 \), where \( \lambda, (a, b), \) and \( (c, d) \) are the scaling parameters of Laplacian distribution, student-t distribution and GPD, respectively. It is seen that the PDF of GPD has narrower pulse width and higher tail values than Laplacian distribution and student-t distribution. Therefore, better performance on sparse representation can be expected with the LSM prior.

III. SPARSE SIGNAL RECOVERY WITH THE LSM PRIOR

In this section, the sparse signal recovery with the LSM prior is derived. Firstly, the MAP estimation based on the expectation-maximization (EM) is presented, which avoids analytical computation of the posterior of the sparse signal but suffers from structural error. Then the EM based variational Bayesian inference using the Laplace approximation [21] is proposed. Unlike the EM based MAP estimation, the EM based variational Bayesian inference using the Laplace estimation is a full Bayesian method and can analytically achieve the posterior of the sparse signal.

A. EM BASED MAP ESTIMATION

The MAP estimation of the sparse signal, \( w \), is expressed as:

\[
\hat{w}_{MAP} = \arg \max_w p(w|y, \lambda, \beta)
\]

which can be solved by the EM algorithm. It is noted that the use of the EM algorithm to solve MAP estimation is not the same as that to solve the MLE. The goal function of EM-MAP is the posterior \( p(w|y, \lambda, \beta) \) and that of EM-MLE is the likelihood \( p(y|w, \lambda, \beta) \).

The EM based MAP estimation is an iterative algorithm consisting the E-step and the M-step.

**E-step:** The Q-function in this step is given in [19]:

\[
Q(w; \hat{w}^{(t)}) = E_{\lambda, \beta|y, \hat{w}^{(t)}} \left[ \log p(w|y, \lambda, \beta) \right]
\]

\[
= E_{\lambda, \beta|y, \hat{w}^{(t)}} \left[ \log p(y|w, \beta) + \log p(w|\lambda) + const \right] \quad (10)
\]

where \( E_{\lambda, \beta|y, \hat{w}^{(t)}} [\cdot] \) denotes the expectation with respect to \( p(\lambda, \beta|y, \hat{w}^{(t)}) \). \( \hat{w}^{(t)} \) represents the sparse signal estimated in the t-th iteration, and \( const \) is a constant independent of \( w \).

**M-step:** The sparse signal is updated by maximizing the Q-function achieved in the E-step:

\[
\hat{w}^{(t+1)} = \arg \max_w Q(w; \hat{w}^{(t)}) \quad (11)
\]
Apparently, calculation of the Q-function is the kernel of the EM algorithm. In order to obtain the Q-function in (10), the posteriors of the scaling parameter, \( p(\lambda, \beta | y, \hat{w}^{(t)}) \), should be derived. Assuming the posterior independence between \( \lambda \) and \( \beta \):

\[
p(\lambda, \beta | y, \hat{w}^{(t)}) = p(\lambda | y, \hat{w}^{(t)}) p(\beta | y, \hat{w}^{(t)}). \tag{12}
\]

Noting that the prior of \( \beta \) is a Gamma distribution being conjugate to the Gaussian distribution, its posterior is also Gamma distributed under the Gaussian noise condition [9] as

\[
p(\beta | y, \hat{w}^{(t)}) = \mathcal{G}\left( \beta | a + \frac{N}{2}, b + \frac{1}{2} \left\| y - \Phi \hat{w}^{(t)} \right\|^2 \right). \tag{13}
\]

Then

\[
E_{\lambda, \beta | y, \hat{w}^{(t)}}[\beta] = \frac{2a + N}{2b + \left\| y - \Phi \hat{w}^{(t)} \right\|^2} \tag{14}
\]

where \( \langle \cdot \rangle \triangleq E_{\lambda, \beta | y, \hat{w}^{(t)}}[\cdot] \). Similarly, the posterior of \( \lambda \) is inverse-Gamma distributed that is conjugate to the Laplacian distribution.

\[
p(\lambda | y, \hat{w}^{(t)}) = \prod_{m=1}^{M} \mathcal{IG}\left( \lambda_m | c + 1, d + \left| \hat{w}_m^{(t)} \right| \right). \tag{15}
\]

Therefore,

\[
\left\{ \frac{1}{\lambda_m} \right\} \triangleq \frac{c + 1}{d + \left| \hat{w}_m^{(t)} \right|}. \tag{16}
\]

Putting (5) and (6) into (10) and ignoring the constant independence of \( w \), the Q-function changes to

\[
Q(w; \hat{w}^{(t)}) = \langle \log p(y | w, \beta) + \log p(w | \lambda) \rangle
= -\frac{\langle \beta \rangle}{2} \| y - \Phi w \|^2 - \sum_{m=1}^{M} \left\{ \frac{1}{\lambda_m} \right\} |w_m| \tag{17}
\]

where \( \langle \beta \rangle \) and \( \{1/\lambda_m\} \) can be derived from (14) and (16), respectively. Then the M-step can be simplified as

\[
\hat{w}^{(t+1)} = \min_w \left[ \frac{\langle \beta \rangle}{2} \| y - \Phi w \|^2 + \sum_{m=1}^{M} \left\{ \frac{1}{\lambda_m} \right\} |w_m| \right]. \tag{18}
\]

The updating process of EM-MAP is equivalent to the re-weighted \( l_1 \) norm minimization method [20].

**B. EM BASED VARIATIONAL BAYESIAN INFERENCE USING THE LAPLACE APPROXIMATION**

Unlike the MAP, the associated posterior of all the latent variables should be derived in the Bayesian inference:

\[
p(w, \lambda, \beta | y) = p(y | w, \beta) p(w | \lambda) p(\beta) p(\lambda) / p(y) \tag{19}
\]

The marginal likelihood of the observation is achieved by integrating the latent variables as

\[
p(y) = \int \int \int p(y | w, \beta) p(w | \lambda) p(\beta) p(\lambda) dw d\lambda d\beta. \tag{20}
\]

Since it is difficult to compute the integral analytically, approximate Bayesian inference methods should be applied. The variational Bayesian inference is a widely applied method, in which the associated posterior is assumed to be factorable, i.e., the latent variables are independent with each other [9]:

\[
p(w, \lambda, \beta | y) \approx q(w, \lambda, \beta) = q(w) q(\lambda) q(\beta). \tag{21}
\]

Based on the EM method, \( q(w) \) is expressed as

\[
\log q(w) = \langle \log p(y | w, \lambda, \beta) \rangle_{q(\lambda q(\beta))} + \text{const}
= \langle \log p(y | w, \beta) p(w | \lambda) \rangle_{q(\lambda q(\beta))} + \text{const} \tag{22}
\]

where \( \langle \cdot \rangle_{q(\lambda q(\beta))} \) represents the expectation with respect to PDF \( q(\lambda) q(\beta) \). For simplification, \( \langle \cdot \rangle_{q(\lambda q(\beta))} \) is expressed as \( \langle \cdot \rangle \) in the following description. It is a challenge to analytically compute the posterior, \( q(w) \), because the Laplacian distribution in (6) is not conjugate to the Gaussian likelihood in (5). Let us consider the Laplace approximation [21] by deriving the second order Taylor expansion of \( \log q(w) \) around the MAP estimation of \( w \),

\[
L(w) = -\langle \log p(y | w, \beta) p(w | \lambda) \rangle
= \langle \beta \rangle^T \Phi^T \Phi + \Lambda \rangle w - \langle \beta \rangle^T \Phi^T y \tag{23}
\]

where \( \Lambda = \text{diag} \left[ \left\{ \frac{1}{\lambda_m} \right\} |w_m| \right] \), and diag[.] denotes a diagonal matrix whose elements is defined in the bracket. Setting \( \nabla_w L(w) \) to zero, we can obtain \( w_{\text{MAP}} \) as

\[
\hat{w}_{\text{MAP}} = \langle \beta \rangle \left( \langle \beta \rangle^T \Phi + \Lambda \right)^{-1} \Phi^T y \tag{25}
\]

Then the Laplace approximation is introduced to approximate \( q(w) \) as

\[
\log q(w) \approx \log q(w_{\text{MAP}}) + \frac{1}{2} (w - w_{\text{MAP}})^T H(w_{\text{MAP}})(w - w_{\text{MAP}}) \tag{26}
\]

which is the second order Taylor expansion of \( \log q(w) \) around the MAP estimation of \( w \). The first order term \( (w - w_{\text{MAP}})^T \nabla_w \log q(w)|_{w=w_{\text{MAP}}} \) equals zero, since \( \nabla_w \log q(w)|_{w=w_{\text{MAP}}} = 0 \). \( H(w_{\text{MAP}}) \) is the Hessian of \( \log q(w) \) evaluated at \( w_{\text{MAP}} \), derived as

\[
H(w_{\text{MAP}}) = \nabla_w^2 \log q(w)|_{w=w_{\text{MAP}}} = -\nabla_w^2 L(w)|_{w=w_{\text{MAP}}} \approx -\langle (\beta) \Phi^T \Phi + \Lambda \rangle \tag{27}
\]
where $\nabla_w^2 L(w) \approx (\beta) \Phi^T \Phi + \Lambda$ is achieved from (24) by regarding $\Lambda$ as independent of $w$. Then the posterior of the sparse signal can be approximately achieved from (26) as

$$q(w) \approx \frac{1}{C} \exp \left\{ -\frac{1}{2} (w - \hat{w}_{MAP})^T \Sigma^{-1} (w - \hat{w}_{MAP}) \right\}$$

(28)

where $\Sigma = (\langle \beta \rangle \Phi^T \Phi + \Lambda)^{-1}$, or $\Sigma = \Lambda^{-1} - \Lambda^{-1} \Phi^T (\langle \beta \rangle^{-1} \Phi + \Lambda^{-1} \Phi) \Phi \Lambda^{-1}$ in the form of the Woodbury matrix identity. $C$ is a normalizing constant given as $C = \int_{-\infty}^{+\infty} \exp \left\{ -\frac{1}{2} (w - \hat{w}_{MAP})^T \Sigma^{-1} (w - \hat{w}_{MAP}) \right\} dw$.

As it resembles the Gaussian distribution, $q(w)$ can be approximated by

$$q(w) \approx \mathcal{N}(w | \mu, \Sigma)$$

(29)

where

$$\mu = \hat{w}_{MAP} = (\beta) \Sigma \Phi^T y.$$  

(30)

Similar to (22), the posterior, $q(\lambda)$, is derived as

$$\log q(\lambda) = \langle \log p(\mathbf{w}, \mathbf{y}, \lambda, \beta) \rangle_{q(w)q(\beta)}$$

$$= \langle \log p(\mathbf{w}|\lambda) \rangle_{q(\mathbf{w}|q(\beta))} + \text{const}$$

$$= -\sum_{m=1}^{M} \log \lambda_m - \sum_{m=1}^{M} \frac{1}{\lambda_m}$$

$$- (c + 1) \sum_{m=1}^{M} \log \lambda_m - d \sum_{m=1}^{M} \frac{1}{\lambda_m} + \text{const}$$

$$= \left[ - (c + 1) - 1 \right] \sum_{m=1}^{M} \log \lambda_m$$

$$- \sum_{m=1}^{M} (d + \langle |w_m| \rangle) \frac{1}{\lambda_m} + \text{const}. $$

(31)

It is seen that the posterior $q(\lambda)$ is also inverse-Gamma distributed

$$q(\lambda) = \prod_{m=1}^{M} \mathcal{I}\mathcal{G}(\lambda_m | \tilde{c}, \tilde{d}_m)$$

(32)

where

$$\tilde{c} = c + 1$$

(33)

$$\tilde{d}_m = d + \langle |w_m| \rangle.$$  

(34)

Also, the posterior of $\beta$, $q(\beta)$, is derived as

$$\log q(\beta) = \langle \log p(\mathbf{w}, \mathbf{y}, \lambda, \beta) \rangle_{q(\mathbf{w|q(\lambda)})}$$

$$= \langle \log p(\mathbf{y}|\mathbf{w}, \beta) \rangle_{q(\mathbf{w}|q(\lambda))} + \text{const}$$

$$= \frac{N}{2} \log \beta - \frac{\beta}{2} \left\| \mathbf{y} - \Phi \mathbf{w} \right\|^2$$

$$+ (a - 1) \log \beta - b \beta + \text{const}$$

$$= \left( a + \frac{N}{2} - 1 \right) \log \beta$$

$$- \frac{1}{2} \left\| \mathbf{y} - \Phi \mathbf{w} \right\|^2 + b \beta + \text{const}$$

(35)

which is the PDF of a Gamma distribution:

$$q(\beta) = \mathcal{G}(\beta | \tilde{a}, \tilde{b})$$

(36)

where

$$\tilde{a} = a + \frac{N}{2}$$

(37)

$$\tilde{b} = b + \frac{1}{2} \left\| \mathbf{y} - \Phi \mathbf{w} \right\|^2.$$  

(38)

With the posteriors shown in (29), (32) and (36), the sparse signal and latent variables can be estimated by iteratively updating the expectations of these posteriors until the required estimating precision is achieved. The updating of expectations is given as follows:

$$\langle \mathbf{w} \rangle = \mu$$

$$= \langle \beta \rangle \langle \Phi^T \Phi \rangle$$

$$+ \text{diag} \left( \frac{1}{\langle \lambda_m \rangle} \right)$$

$$\langle \beta \rangle_y = \frac{\tilde{c} + 1}{\tilde{d} + \langle |w_m| \rangle}$$

$$\langle \beta \rangle_y = \frac{\tilde{a}}{\tilde{b}}$$

(39)

(40)

(41)

In order to compute the above updating formulas, the moments of $\mathbf{w}$, including $\langle |w_m| \rangle$ and $\langle \| y - \Phi \mathbf{w} \|^2 \rangle$, should be derived. For $\langle \| y - \Phi \mathbf{w} \|^2 \rangle$, it is given in [9]:

$$\langle \| y - \Phi \mathbf{w} \|^2 \rangle = \| y - \Phi \mu \|^2 + tr \left( \Sigma \Phi^T \Phi \right).$$

(42)

In [22], $\langle |w_m| \rangle$ is shown to be the absolute first moment of the Gaussian distribution in (29) as

$$\langle |w_m| \rangle = \sqrt{\frac{2 \Sigma_{w_m}}{\pi}} \text{erf} \left( \frac{\mu_m^2}{\Sigma_{w_m}} \right)$$

$$= \sqrt{\frac{2 \Sigma_{w_m}}{\pi}} \exp \left( -\frac{\mu_m^2}{\Sigma_{w_m}} \right)$$

(43)

where $\text{erf}(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{2n+1} \pi$ is the Kummererâ€™s confluent hypergeometric functions, $x^{(n)}$ is the rising factorial: $x^{(n)} = x(x+1)(x+2)\cdots(x+n-1)$ and erf($x$) is $2/\sqrt{\pi} \int_0^x e^{-t^2} dt$ is the error function.

In summary, the sparse signal $\mathbf{w}$ in each iteration of the algorithm is updated by (39) with the estimated scaling parameter $\lambda$ and the inverse noise variance $\beta$. Then $1/\langle \lambda_m \rangle$ and $\beta$ are sequentially updated by (40) and (41), respectively.
In addition, the difference between EM-MAP and the proposed approximate variational EM based Bayesian inference is discussed. For EM-MAP, its updating formula can be derived from (18) as:

$$\hat{w}^{(t+1)} = \langle \beta \rangle \Phi^T \Phi + \text{diag} \left( \frac{1}{\lambda_m} \right) \Phi^T y.$$  

(44)

Comparing (44) with (39), it is found that the difference lies in computing the elements of the diagonal matrix in the variance matrix. The absolute of the previous estimation result, $|\hat{w}_m^{(t)}|$, is used to update the variance matrix in (44), while in (39), it is updated by the first absolute moment of the posterior, $|\langle |w_m| \rangle|$, which is achieved by (43). In other words, the proposed EM based variational Bayesian inference using the Laplace approximation utilizes more statistical information of the sparse signal than the EM-MAP. Therefore, performance improvement on sparse signal recovery can be expected.

C. EXTENSION TO GENERAL PRIORS

The proposed variational Bayesian inference based on the Laplace approximation leads to full Bayesian inference for any kinds of prior. That is, the posterior of sparse signals with any kinds of prior can be approximated by the Laplace approximation and used in the variational Bayesian inference. For example, let us consider the derivation of the variational Bayesian inference based on the Laplace approximation with the GPD prior. Then the prior of the sparse signal can be expressed as the marginal likelihood in (8), and the MAP estimation in (18) is changed to

$$\hat{w}^{(t+1)} = \arg \min_{w} \left[ \langle \beta \rangle \frac{1}{2} ||y - \Phi w||^2 ight.$$ 

$$+ (c+1) \sum_{m=1}^{M} \log (|w_m| + d) \right].$$  

(45)

With similar derivations in (23)-(25), the following updating expression can be obtained:

$$\hat{w}_{\text{MAP}} = \langle \beta \rangle \left( \langle \beta \rangle \Phi^T \Phi + \Lambda_{\text{GPD}} \right)^{-1} \Phi^T y$$  

(46)

where $\Lambda_{\text{GPD}} = \text{diag} \left( \frac{1}{|w_m| + (|w_m| + d)} \right)$. Then the posterior of the sparse signal can be approximated by the Laplace approximation in (26), and is found to be Gaussian distributed approximately:

$$q_{\text{GPD}}(w) \approx \mathcal{N} \left( w | \mu_{\text{GPD}}, \Sigma_{\text{GPD}} \right)$$  

(47)

where $\mu_{\text{GPD}} = \langle \beta \rangle \Sigma_{\text{GPD}} \Phi^T y$, $\Sigma_{\text{GPD}} = \left( \langle \beta \rangle \Phi^T \Phi + \Lambda_{\text{GPD}} \right)^{-1}$.

Comparing (47) with (29), it is seen that the only difference is the diagonal matrix in the variance matrix. Actually, the posterior $q_{\text{GPD}}(w)$ in (47) is equivalent to the posterior $q(w)$ in (29):

$$\Lambda = \text{diag} \left[ \frac{1}{\lambda_m} \right]$$

$$= \text{diag} \left[ \frac{1}{|w_m|} \right] \left( \frac{c+1}{d + |\langle |w_m| \rangle|} \right)$$  

(48)

where $\Lambda$ is derived in (40). Therefore, with the utilization of the proposed approximate variational Bayesian inference algorithm, the same updating expressions of the sparse signal can be achieved with LSM prior and GPD prior. It should be noted that if the MAP estimation can be derived for a specific prior, its approximate full Bayesian inference can be accomplished by the proposed algorithm, in which the posterior of sparse signals is estimated by the second Taylor expression around its MAP estimation.

D. EXTENSION TO COMPLEX DATA

The proposed approximate variational Bayesian inference with LSM prior is capable of recovering the complex sparse signal with some adjustments. For the complex sparse signal recovery, the additive noise, $n$, in (1) is complex Gaussian distributed, which changes the likelihood of the observation, $y$, in (5) to

$$p(y|w, \beta) = \pi^{-N} |\beta|^N \exp \left( -\beta^2 ||y - \Phi w||^2 \right).$$  

(49)

Accordingly, the objective function $L(w)$ in (23) is changed to

$$L(w) = \langle \beta \rangle ||y - \Phi w||^2 + \sum_{m=1}^{M} \left( \frac{1}{\lambda_m} \right) |w_m| + \text{const.}$$  

(50)

Noting that $w$ is a complex variable, the first derivative of $L(w)$ with respect to $w^*$ represents the fastest descending direction of $L(w)$, which is derived as

$$\nabla_w L(w) = \left( \langle \beta \rangle \Phi^H \Phi + \frac{1}{2} \Lambda \right) w - \langle \beta \rangle \Phi^H y.$$  

(51)

From (51), the MAP estimation of $w$ is obtained as

$$\hat{w}_{\text{MAP}} = \langle \beta \rangle \left( \langle \beta \rangle \Phi^H \Phi + \frac{1}{2} \Lambda \right)^{-1} \Phi^H y$$  

(52)

where $\Lambda = \text{diag} \left[ \frac{1}{\lambda_m} \right]$. Putting it into (28), the approximate posterior of $w$ is derived as

$$q(w) \approx \mathcal{CN} \left( w | \mu_C, \Sigma_C \right)$$  

(53)

where $\mathcal{CN}(w | \mu_C, \Sigma_C)$ denotes a complex Gaussian distribution with an expectation of $\mu_C = \langle \beta \rangle \Sigma_C \Phi^H y$ and a variance matrix of $\Sigma_C = \left( \langle \beta \rangle \Phi^H \Phi + \frac{1}{2} \Lambda \right)^{-1}$. Comparing (53) with (29), it is found that the complex posterior is obtained by simply replacing the transpose operator, $(\cdot)^T$, with the conjugate transpose operator, $(\cdot)^H$, in the real posterior in (29).

\footnote{The complex Gaussian distribution is defined as $p(x | \mu, \Sigma) = \frac{1}{\pi^n \det(\Sigma)^{1/2}} \exp \left[ -\frac{1}{2} \left( x - \mu \right)^H \Sigma^{-1} \left( x - \mu \right) \right]$, where $x \in \mathbb{C}^n$.}
and changing the coefficient of the diagonal matrix, Λ, in Σ to 1/2.

IV. EXPERIMENTS

In this section, simulated experimental results are analyzed to validate the effectiveness of the proposed EM based variational Bayesian inference using the Laplace approximation (known as EM-VB). Then one practical application of the EM-VB, i.e., Inverse Synthetic Aperture Radar (ISAR) imaging, is implemented to further testify its performance.

A. SIMULATED DATA

EM-VB is compared with OMP [6], [7], SBL [10], [11], BCSL [15], [24] and the EM based MAP estimation with the LSM prior (known as EM-MAP) presented in section III-A. All the parameters of these algorithms are suitably adjusted to achieve their best performance. The experimental setups in the following experiments are the same as those in [15].

The first experiment directly compares the sparse signals recovered by the aforementioned five algorithms. In this experiment, the number of observations and sparse coefficients are N = 100 and M = 512, respectively. The measurement matrix Φ is chosen as a uniform spherical ensemble, where each column is uniformly distributed on the unit sphere. Similar results are also obtained by other measurement matrices such as partial Fourier and uniform random projection (URP) ensembles [15]. Only the uniform spherical ensemble is considered in this paper. The sparse signal contains K = 20 randomly distributed nonzero coefficients with identical and standard normally distributed magnitudes, respectively. In addition, the zero-mean Gaussian white noise is added to the observation to simulate the noisy environment, and the noise level is controlled by the standard deviation. In this experiment, it is set as 0.09 to achieve the Signal to Noise Ratio (SNR) of 10 dB. Fig. 3 shows the recovery of the identical and standard normally distributed spikes by these five algorithms, and the corresponding reconstruction errors (∥w − w∥2/∥w∥2) are given in Table 1. It is seen that the proposed EM-VB accurately recovers the sparse signal with little noise and gives the least reconstruction error compared to other algorithms. In contrast, EM-MAP obtains the most distorted sparse signals and the highest reconstruction error among the five algorithms. It indicates that the EM-MAP produces structural error and is a “poor man’s Bayesian Inference” [9]. The EM-VB, with the processing of fully Bayesian inference, avoids the structural error. It not only suppresses the noise, but also precisely recovers the magnitude of the spikes.

Next, we compare the performance of five algorithms under the noise-free condition. In this case, the number of sparse coefficients and non-zero elements in sparse signal is M = 512 and K = 20, respectively. The number of observations, N, is varied from 40 to 120 in steps of 20, so as to indicate the reconstruction error and 95% confidence interval versus N. Averaged reconstruction error of 100 trials is collected, in which a trial means a random assignment of the sparse coefficients and compressive measurement matrix. Fig. 4(a) and Fig. 4(b) show the curves of the reconstruction error versus N from different algorithms for sparse signals with identical and normal distributed spikes, respectively, and Fig. 4(c), (d) give the respective half 95% confidence interval. It is seen that among the five algorithms, EM-VB achieves the least reconstruction errors and the narrowest confidence interval. Its performance is more remarkable for the identical spikes compared with the normal distributed spikes. In addition, it should be noted that all error curves descend similarly as the increase of N except that of EM-MAP, whose descending trend is much flatter. It indicates that EM-MAP does not belong to the Bayesian sparse signal recovery algorithms.

In the next experiment, we mainly compare the performances of different algorithms versus the number of observations, N, in noisy environment with SNR = 8 dB and 16 dB. N is varied from 80 to 140 in steps of 10, and the rest experimental setups are the same as the first experiment. Fig. 5 presents the reconstruction error curves and half 95% confidence interval versus N for sparse signals with identical and standard normally distributed spikes, in which all the reconstruction errors are averaged with 100 trials, and each trial contains random assignments of the sparse coefficients, compressive measurement matrix and noise vector. It is seen that the EM-VB algorithm performs slightly better than EM-MAP algorithm when SNR = 8 dB. When SNR = 16 dB, however, the gap between the curves of EM-VB and EM-MAP becomes

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>OMP</th>
<th>SBL</th>
<th>BCSL</th>
<th>EM-MAP</th>
<th>EM-VB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical spikes</td>
<td>0.3001</td>
<td>0.2790</td>
<td>0.2629</td>
<td>0.6030</td>
<td>0.1231</td>
</tr>
<tr>
<td>Standard normally distributed spikes</td>
<td>0.2432</td>
<td>0.2188</td>
<td>0.1997</td>
<td>0.5945</td>
<td>0.1443</td>
</tr>
</tbody>
</table>
obvious in both Fig. 5(a) and Fig. 5(b). It means that although EM-MAP achieves a similar performance to EM-VB under low SNR condition, its performance is highly affected by the structural error when SNR is high. In contrast, this problem is avoided by EM-VB. This comparison reveals the performance improvement introduced by the Laplace approximation on the posterior used in EM-VB. Comparing the EM-VB with SBL and BCL, it is seen that EM-VB achieves the least reconstruction error in most situations, except when SNR = 16 dB for standard normally distributed spikes, as shown by the red dashed curves in Fig. 5(b). The more obvious performance improvement by EM-VB for SNR = 8 dB than 16 dB indicates that it is more robust to noise than SBL and BCSL. Additionally, the proposed EM-VB also achieves the least confidence interval in most cases, which further validates its superior stability.

The next experiment mainly compares the performances of five algorithms with respect to noise level. The parameters for this experiment are $K = 20$, $M = 512$, SNR = 4 dB to 16 dB in steps of 2 dB, and $N = 100$ and 160, respectively. Fig. 6 shows the curves of averaged reconstruction errors and half 95% confidence interval of 100 trials versus SNR for identical and standard normally distributed spike signals, respectively. Also, each trial contains random assignments of the sparse coefficients, compressive measurement matrix and noise vector. It can be seen that the proposed EM-VB achieves the least reconstruction error and confidence interval in most cases. Although EM-MAP performs slightly
better than EM-VB when SNR is low, e.g. SNR < 8 dB, its performance deteriorates dramatically as N decreases. When N = 100, the reconstruction error curve of EM-MAP becomes much flatter than those for the other algorithms, which indicates the decrease of N aggravates the reconstruction error achieved by EM-MAP. In contrast, EM-VB maintains its superiority under any conditions, which further validates its effectiveness. In addition, comparing Fig. 6(a) with Fig. 6(b), it is interesting to see that the difference between the error curves with N = 100 and those with N = 160 for identical spikes is much higher than that for standard normally distributed spikes, which indicates that the reconstruction of the former is more sensitive to the number of observations than that of standard normally distributed spike signal.

Last but not least, the performances of five algorithms under different number of sparse coefficients, K, are compared. The parameters for this comparison are N = 160, M = 512, K = 5 to 30 in steps of 5, and SNR = 8 dB and 16 dB, respectively. The averaged reconstruction errors and half 95% confidence interval of 100 trials, with each trial containing random assignments of the sparse coefficients, compressive measurement matrix and noise vector, for identical and standard normally distributed spike signals are given in Fig. 7. For the identical spikes, EM-VB achieves the best performance among the five algorithms when K > 10. For the standard normally distributed spike signal and SNR = 16 dB, the reconstruction error curves of EM-VB, BCS and BCSL are close and much lower.
FIGURE 6. Reconstruction error (the first row) and half 95% confidence interval (the second row) versus SNR for identical (the first column) and standard normally distributed (the second column) spike signals with $N = 100$ and 160.

than that of EM-MAP. When SNR is decreased to 8 dB, the performance of EM-MAP becomes comparable as that of EM-VB, and is better than those of BCS and BCSL. Additionally, EM-VB achieves the least confidence interval in most cases. It is seen that because of full Bayesian inference used in EM-VB, it achieves both the robustness to noise from EM-MAP and the ability of avoiding structural error from BCS and BCSL.

In summary, experimental results under various conditions in terms of the number of observations, noise level, and the number of non-zero elements of sparse signal for the identical and standard normally distributed spike signals validate that the proposed EM-VB outperforms EM-MAP, BCS and BCSL in terms of recovery precision and the robustness to noise. It should be mentioned that although only results with two kinds of sparse signals are presented in this paper, numerous experiments on other kinds of sparse signals, such as the Laplacian distributed and student-t distributed spike signals have also been conducted. Similar results have been obtained to indicate the performance superiority of EM-VB. Therefore, we are confident to conclude that it achieves state-of-art performance for sparse signal recovery.

B. MEASURED DATA
The proposed EM-VB as been used for practical application to further validate its effectiveness. With the ability
Reconstruction error (the first row) and half 95% confidence interval (the second row) versus the number of non-zero elements of sparse signal for identical (the first column) and standard normally distributed (the second column) spikes signals with SNR = 8 dB and 16 dB.

FIGURE 7

In obtaining high resolution images of the moving targets (airplanes, vessels, etc.) in all-day and all-weather environment, Inverse Synthetic Aperture Radar (ISAR) imaging is of great significance for both civil and military purposes [25]. It is often that only sparse aperture (SA) signal is available due to nonideal system operation and/or strong interferences, which challenges the traditional ISAR imaging algorithms [26]. Since the ISAR targets generally exhibit sparse characteristic in the radar reflection field, ISAR imaging for sparse aperture data can be accomplished within the sparse signal recovery framework [27]. In this experiment, a set of real data measured by a C-band radar with a center frequency of 5.52 GHz and a bandwidth of 400 MHz is utilized to imaging An-26 (a twin-engined turboprop medium sized airfreighter), which is shown in Fig. 8(a). The full radar echo contains 512 pulses from which 128 pulses are randomly sampled, as given in Fig. 8(b), to simulate the sparse aperture data. The radar images and their image entropy\(^3\) obtained by the traditional Range Doppler algorithm (RD) [28], OMP, SBL, BCSL, EM-MAP and EM-VB, are given in Fig. 8 and Table 2, respectively. It is seen that the proposed EM-VB obtains the clearest image as well as the least image entropy.

\(^3\)Image entropy is generally used to measure the image quality in ISAR imaging, and is defined as [29]:

\[
E = - \sum_i \sum_j |x(i,j)|^2 / P \log |x(i,j)|^2 / P,
\]

where \(x\) denotes the radar image and \(P = \sum_i \sum_j |x(i,j)|^2\) denotes the image energy. Lower image entropy indicates better focused radar image.
Laplacian distributions used to model the sparse coefficients the LSM prior. Firstly, the distinct scaling parameters of the utilized to accomplish full sparse Bayesian inference with ness than the widely applied GSM prior. Some strategies are sparse signal, which performs better in representing sparse-approximation. The LSM prior is introduced to model the using the variational Bayesian inference based on the Laplace This paper proposes a new sparse signal recovery algorithm ISAR imaging.

V. CONCLUSION
This paper proposes a new sparse signal recovery algorithm using the variational Bayesian inference based on the Laplace approximation. The LSM prior is introduced to model the sparse signal, which performs better in representing sparseness than the widely applied GSM prior. Some strategies are utilized to accomplish full sparse Bayesian inference with the LSM prior. Firstly, the distinct scaling parameters of the Laplacian distributions used to model the sparse coefficients are assumed to be inverse-Gamma distributed, which is conjugate to the Laplacian distribution. Then the posterior of the sparse signal is assumed by the Laplace approximation and found to be Gaussian distributed with the expectation being the MAP result of the sparse signal. Following that, the EM based variational Bayesian inference with the LSM prior is derived. Compared with the EM-MAP, the proposed EM-VB algorithm obtains both the estimation of the sparse signal and its second order statistical information. Experimental results based on both simulated and measured data validate the state-of-art performance of the proposed algorithm.

APPENDIX
DERIVATION OF THE MARGINAL LIKELIHOOD IN (8)
The marginal likelihood shown in (8) can be derived as

\[
p(w_m; c, d) = \int_0^{+\infty} p(w_m | \lambda_m) p(\lambda_m; c, d) \, d\lambda_m
\]

\[
= \int_0^{+\infty} \frac{1}{2\lambda_m} \exp\left(-\frac{|w_m|}{\lambda_m}\right) \frac{d^c}{\Gamma(c)} \lambda_m^{-c-1} \exp\left(-\frac{d}{\lambda_m}\right) \, d\lambda_m
\]

\[
= \frac{d^c}{2\Gamma(c)} \int_0^{+\infty} \lambda_m^{-c-2} \exp\left(-\frac{|w_m| + d}{\lambda_m}\right) \, d\lambda_m
\]

(A.1)

Let \( u = (|w_m| + d)/\lambda_m\), then \( d\lambda_m = u^{-2} (|w_m| + d) \, du \).

Putting them into (53), we obtain

\[
p(w_m; c, d) = \frac{d^c}{2\Gamma(c)} \int_0^{+\infty} \lambda_m^{-c-2} \exp\left(-\frac{|w_m| + d}{\lambda_m}\right) \, d\lambda_m
\]

\[
= \frac{d^c}{2\Gamma(c)} \int_0^{+\infty} [(|w_m| + d) u^{-1}]^{-c-2} \times \exp(-u) \, u^{-2} (|w_m| + d) \, du
\]

\[
= \frac{d^c}{2(|w_m| + d)^{c+1}} \Gamma(c+1) \int_0^{+\infty} u^c \exp(-u) \, du
\]

\[
= \frac{c d^c}{2(|w_m| + d)^{c+1}} \Gamma(c+1)
\]

(A.2)

which indicates its superior performance for sparse aperture ISAR imaging.

TABLE 2. Image entropy of the radar images in Fig. 8.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>RD</th>
<th>OMP</th>
<th>SBL</th>
<th>BCSL</th>
<th>EM-MAP</th>
<th>EM-VB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image entropy</td>
<td>8.8265</td>
<td>6.4053</td>
<td>5.8453</td>
<td>5.5599</td>
<td>5.5838</td>
<td>4.4324</td>
</tr>
</tbody>
</table>

FIGURE 8. Radar images of An-26 obtained by different algorithms. (a) An-26; (b) 128 randomly sampled range profiles; (c), (d), (e), (f), (g) and (h) Radar images obtained by RD, OMP, SBL, BCSL, EM-MAP and EM-VB, respectively.

REFERENCES


SHUANGHUI ZHANG received the B.S. and Ph.D. degrees from the National University of Defense Technology (NUDT), in 2011 and 2016, respectively. He is currently a visiting Ph.D. student of Nanyang Technological University in 2015.

He is currently a Lecturer with the College of Electrical Science in NUDT. His research interests lie in compressive sensing, sparse signal recovery techniques, Bayesian inference, and their applications in radar signal processing.

YONGXIONG LI (M’13) received the B.S. and Ph.D. degrees from the National University of Defense Technology, in 1999 and 2004, respectively.

Since 2004, he has been with NUDT, where he is currently a Professor with the College of Electrical Science and Engineering, conducting research on radar target recognition, time-frequency analysis, and micro-motions. He was with the Imperial College London as an Academic Visitor in 2008.

XIANG LI (M’09) received the B.S. degree from Xidian University, Xi’an, in 1989, and the M.S. and Ph.D. degrees from the National University of Defense Technology in 1995 and 1998, respectively.

He is currently a Professor and performs research on auto target recognition, signal detection, and nonlinear signal processing.

GUOAN BI (SM’89) received the B.Sc. degree in radio communications from the Dalian University of Technology, China, 1982, and the M.Sc. degree in telecommunication systems, and the Ph.D. degree in electronics systems from Essex University, U.K., in 1985 and 1988, respectively.

Since 1991, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include DSP algorithms and hardware structures, and signal processing for various applications including sonar, radar, and communications.

S. ZHANG et al.: Variational Bayesian Sparse Signal Recovery With LSM Prior